Interactive Object Recognition with Sensor Fusion

László Czúni
Image Processing Laboratory
University of Pannonia
Veszprém, Hungary
Email: czuni@almos.uni-pannon.hu

Metwally Rashad
Image Processing Laboratory
University of Pannonia
Veszprém, Hungary
Email: metwally.rashad@virt.uni-pannon.hu

***Abstract*—A new approach for lightweight video-based object recognition is introduced where a user moves a camera around a target object of interest. The extraction of image features and the retrieval algorithm is running in a lightweight mobile computer (tablet or phone). We apply a view based model of the objects and the matching of the query and candidate images is based on compact image descriptors coupled with relative orientation.**

I. Introduction

Object recognition can go beyond simple detection and requires cognitive processes to recognize, interpret, and appropriately respond to objects. Because survival can depend on the ability to perform these operations quickly and accurately, it follows that the evolution of cognitive processes necessary for object recognition should be widespread. It does not necessarily follow, however, that the nature of these processes is the same across all species. Also there are different approaches in artificial systems for object recognition depending on the task, the type of sensors, and the constraints (e.g. complexity, memory size), while nowadays the use of mobile devices, where computational burden should be kept low, means new challenges. While optical recognition has many problems in general such as scaling, illumination changes, partial occlusion, and background clutter, in case of capturing 3D objects with mobile devices viewpoint variation and image noise (e.g. motion blur due to hand shaking in poor lighting conditions) can decrease the recognition rate tremendously. Numerous recognition algorithms have been developed, most of them apply single image-based recognition, taking only one image as input. However, object recognition from a single view may fail when there is much similarity among the captured image of objects or when the background clutter or partial occlusion masks distinctive features. Video based approaches can use more views but suffer from the increased amount of information resulting in a need for efficient lightweight but robust techniques. In our paper we discuss a new approach for video-based object recognition where a user moves a mobile camera around a target object of interest, while keeping the object roughly in the center of the viewfinder. The capturing of images and orientation data is done at multiple viewpoints in an arbitrary manner with variations in orientation. The extraction of image features and the retrieval algorithm is running in the lightweight client (an Android based mobile phone or tablet) without a need for client-server communication with server side processing.

The models, used for the recognition of objects, are generated by previous recordings and are built from images taken from different viewing directions. The orientation of the camera is also recorded besides the image descriptors generated from the color images. In our experiments we use the compact color and edge directivity descriptor (CEDD) as feature vector and use the Tanimoto distance for matching.

II. Related Works

In the recent years there is an increasing number of papers dealing with object modeling and object recognition topics, we just mention a few related to our work. In [1] recognition was achieved from the video sequences by employing a multiple hypothesis approach. Appearance similarity, and pose transition smoothness constraints were used to estimate the probability of the measurement being generated from a certain model hypothesis at each time instant. A smooth gradient direction feature was used to represent the appearance of object while the pose of the object at each time instant was modeled as a von Mises-Fisher distribution. Recognition was achieved by choosing the hypothesis set that has accumulated the maximum evidence at the end of the sequence. Unfortunately, the testing of the method was carried out on four objects only. In [14] view-based object recognition based on human perception is introduced by suggests that the human brain represents objects as a series of interconnected views and proposed a system which learns such representations of objects through the process of feature tracking. The concept of key-frames which are acquired from the visual input allows for natural characterization of the visual complexity in the input. In [2] authors created object models with the help of SIFT points which are tracked from frame to frame. Video matching is based on the comparison of every query frame with all components of all models. While the accuracy was about 83% in case of 25 objects, no information about the complexity is given. In [4] in addition to the camera they used the accelerometer and the magnetic sensor to recognize the landscape. Clustered SURF (Speeded Up Robust Features) features were quantized using a vocabulary of visual words, learnt by k-means. For tracking objects the FAST corner detector was combined with sensor tracking. Because of the small storage capacity of the mobile device a server-side service was needed to store the large number of images. In [3] the problem of searching in large databases with mobile devices is attacked. The paper focuses on indexing (with bag of hash bits) and applies saliency based segmentation. It also states that drastic change in camera perspective and/or lighting, too small image/object size, non-rigid objects, insufficient (or non-discriminative) local features can cause serious problems in retrieval. In [5] we showed that CEDD is quite tolerant for different noises and can be computed in today mobile platforms and used for object recognition. Now, we extend

978-1-4673-8129-1/15/$31.00 ©2015 IEEE

this previous framework to use several views and also include the orientation sensor to get better performance.
The proposed new method results in better recognition rate since the multiple view increases the confidence measure of the match.

III. Proposed Method

One of the most intriguing aspects of object recognition is our ability to identify objects across changes in viewpoint. In particular, depth rotations of an object can drastically change the 2-D information. Two general classes have been proposed to explain how recognize objects when seen from novel views using a single camera. The classes differ principally in terms of how the shapes, features, and structure of objects are represented, and what processes are involved in object recognition. In object centered representations object features describe the 3D structure or volume of the object. Classical structure from motion methods (e.g. [6]) can be considered as such solutions. The main disadvantage of these methods is that they require simultaneous calibration of cameras and 3D reconstruction far from being real-time. In case of view centered representations the outlook of the object is modeled from different viewpoints. There is no effort taken to reconstruct the (2D or 3D) structure of the object rather information is collected and organized such a way that can be easily used for recognition. We followed the second approach and the viewpoint was estimated using the orientation sensors of the camera (mobile device).

A. Image Feature Extraction and Comparison

We do not attempt to give a review on image feature extraction in our paper just list some possible methods we thought would serve as the basis of a robust recognition engine. In our previous tests [5] we investigated the following three types of descriptors: MPEG-7 based methods [8] (MPEG7_CLD, MPEG7_EHD, MPEG7_SCD, MPEG7_Fusion); Local feature based methods (SURF, SURFVW [7], SIFT [9]); Compact Composite Descriptors [7] (CompactCEDD, CEDD, CompactFCTH, FCTH, JCD, CCD Fusion, CompactVW). Unfortunately, the SIFT based method ran extremely slow (about two orders slower than compact descriptors) in initial tests compared to others and its performance was not better than the average of all. Even it seemed to be very sensitive to motion blur so it was neglected in our further experiments. Please note that although there are several much faster local descriptors [10] than SIFT, the selection of the most appropriate one is out of focus of this paper.
The selected CEDD descriptor, what was found quite robust in previous works, combines color and texture information of a rectangular region in histograms in a vector of length 144. Texture information of image blocks is modeled by classifying them into six classes: non-edge, vertical, horizontal, 45-degree diagonal, 135-degree diagonal and nondirectional edges. Each class is described by 24 bin color histogram based on fuzzy color selection. The process of generating the CEDD descriptor is described in the flowchart Fig. 1. For more details on CEDD please refer to paper [7]. According to previous tests the similarity of two CEDD descriptor vectors are efficiently given by the Tanimoto coefficient [7]. Let q_i be the descriptor of the ith frame from the query and c_j be the descriptor of the jth

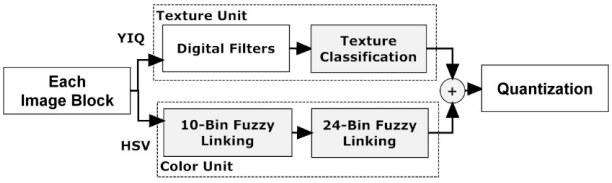

Fig. 1. CEDD Flowchart

frame of a candidate. The Tanimoto Coefficient is then:

$$T(q_i, c_j) = \frac{q_i^T c_j}{q_i^T q_i + c_j^T c_j - q_i^T c_j} \quad (1)$$

Even if the sole of the object is fixed, the relative orientation of the camera (compared to the object) can be changed from time to time and thus the rotation of the camera can be described by pan, tilt, and roll. While we can get rid of the problem of different pan and tilt settings if object tracking is applied (see in later Section) camera roll should be handled differently. CEDD is not rotation invariant but with the modification of the Tanimoto distance (T) rough rotation invariance could be achieved.

B. Model Generation and Retrieval

In our model we have not only one but several CEDD descriptors of the objects extracted from different viewing directions (see Fig. 2 for illustration). In each case the object is located in the center of the image while the elevation and azimuth can be varied due to camera tilt, pan, and translation. Each descriptor is coupled with the orientation data giving the elevation and azimuth in degree measured with the digital compass and acceleration sensors. Azimuth angle should be considered as a relative value since the object can be rotated between two queries, that is we need an azimuth matching mechanism (built into a modified similarity function in the next Section). To reduce the database size several visually similar frames can be removed from the database. Let c_i and c_j be two consecutive frames taken at different azimuths. If the difference in $T(c_i, c_j)$ is below threshold Th then c_j is simply deleted.

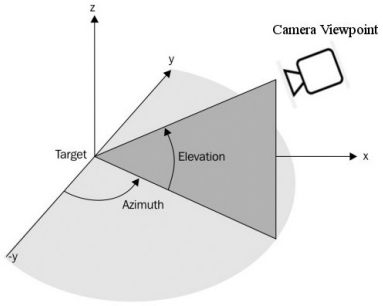

Fig. 2. Model generation setup

In our typical use-case an object is placed on a surface (e.g. table) and the observer moves its camera around it keeping the object roughly in the center. The image can contain several objects so the target object should be segmented from its background. This separation can be easily carried out by

setting a target rectangle manually in the first frame then applying tracking such as Camshift [13] with low complexity. We also implemented automatic segmentation by Grabcut [12] giving nice contours instead of rectangles. The disadvantage of the later is the running time which is about 5 sec in our test platform (specified later). Without using temporal or orientational information one may use several frames from the query video to compute the average Tanimoto Coefficient resulting in complexity $O(N_c \cdot N_f^q \cdot N_f^c)$ where N_c is the number of candidate objects, N_f^q is the number of frames in query and N_f^c is the number of frames in candidates (referring to object model size). Contrary, we show that testing only one frame from the query against all model frames then using the known relative orientation information for the other frames results in much lower complexity but similar hit rate. That is we defined the following similarity function:

$$T^{multi-sensor}(q,c) = \frac{\min_j T(q_i,c_j) + \sum_{\forall k,k\neq i} T(q_k, c_{\alpha(k)})}{N_f^q} \quad (2)$$

where i is randomly selected (in our current implementation) and $\alpha(k)$ means the frame which is at the same (or very close) relative orientation in the candidate model to j as k to i in the query. The complexity of the multi-sensor method can be described as: $O(N_c \cdot (N_f^c + (2 \cdot (N_f^q - 1))))$. Since there is no guarantee that we find a frame at the exact relative position in the candidate we used the best matching of the left and right neighbors in the closest available orientations explaining the multiplication by 2 in the above complexity. Please note, that in this case the search heavily depends on the randomly selected frame (i) used for orientation estimation. In future we plan to use a reduced information set from all available query frames.

IV. Experimental Results

A. Datasets

The model dataset includes 16 objects (fully 3D-shaped and some of them with somehow similar appearance), like some types of cars, headset, books, coffee cups, stapler, plastic bags, computer mouse, some types of pens. Between 44-73 views per object were captured from the same elevation but from different viewing angles (azimuth) leading to approximately 900 images. Objects were centered and a bounding box was manually defined for each image. Image sizes and side ratios varied a lot as shown in Fig. 3. As we can see the object size, shape, color, contrast can vary from view to view. Some view of the same object can be very different from the other (see f.e. the green pencil or the matchbox). The background of the objects were only roughly uniform and the surface of objects was sometimes glossy. We used the built-in accelerometer and compass sensors to measure the orientation of the camera for each view.

The query dataset is composed of 10 randomly selected images of each object strongly distorted by motion blur and additive Gaussian noise. Some examples of the queries are shown in Fig. 4. We think that while the number of objects is not high the very different appearances and heavy distortions make our test somehow realistic.

To reduce the database size we applied a frame selection method as described before. Fig. 5 contains the number of

Fig. 3. Test object examples in increasing ID order. (Only 3 samples per object are shown.)

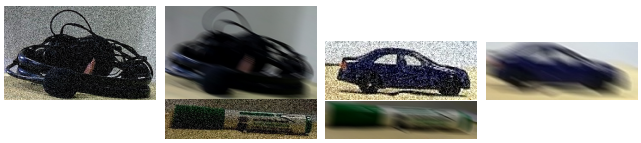

Fig. 4. Noisy and blurred query examples.

frames in case of each object category at different Th threshold settings. The smallest number at threshold 20 was found in case of the white-green bus (7 samples remaining) while the largest number for the green pen (20 samples remaining). This can be easily reasoned by the simple visual examination of the objects.

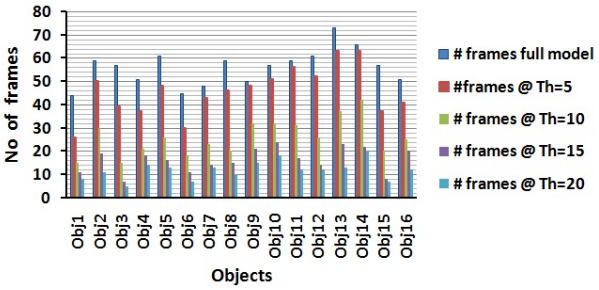

Fig. 5. The number of image samples before and after frame reduction.

B. Retrieval Performance

The purpose of the tests were to see the hit rate and the running time of the multi-sensor method compared to the method when all frames of the query are used for retrieval ("only image" method). The effect of applying different Th-s and N_f^q-s is explored. Since CEDD is very compact memory usage is simply out of interest. Implementations were tested on a Samsung SM-T311 tablet equipped with Android 4.2.2 Jelly Bean, 1 GB RAM, and ARM Cortex A9 Dual-Core 1.5 GHz Processor. We tested different number of query images ($N_f^q = 1,...8$) at random orientations. There are two graphs illustrating the hit rate vs. the number of frames in the query. As Fig. 6 shows for motion blurred images retrieval performance is greatly affected by the model size and as N_f^q goes from 1 to 8 the hit rate increases about 5%. The strong additive noise resulted lower values (see Fig. 7), especially when model size was reduced by $Th = 20$. But the same can be seen: model size greatly influences hit rate and multiple query frames can increase the result with more than 10%.

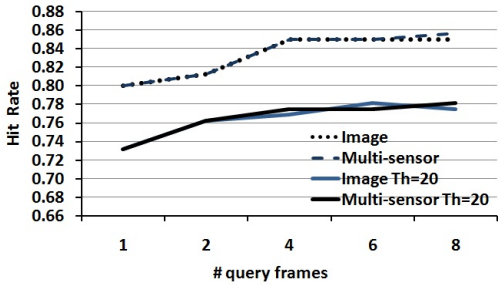

Fig. 6. Average hit rate for strong motion blur.

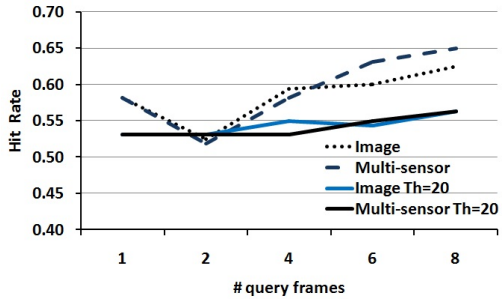

Fig. 7. Average hit rate for strong additive Gaussian noise.

In our tests we ran full search but decision trees can even decrease the running time which might be interesting if the number of objects is much higher. Fig 8 illustrates the average running time (based on 10 queries) for the different retrieval methods. (Please note that the extraction of the CEDD descriptors, which is about 0.4 sec, is not included in these data.) It is clearly visible that as the number of query frames is increasing the advantage of the multi-sensor method is growing while giving practically the same retrieval rate. It means that using the multiple-sensor method at $N_f^q = 8$ we get the best performance at the running time of $N_f^q = 2$ of the only-images approach.

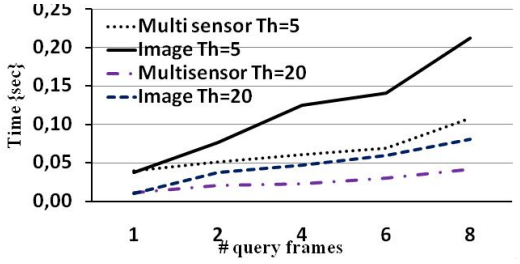

Fig. 8. Average running time for linear image search with different approaches and model size.

V. Conclusions and Future Works

We proposed a multi-image recognition approach with a compact global image descriptor coupled with orientation data. This way the descriptor size and the number of matching steps can be kept low. In tests using over 900 different view images of 16 test objects under different image distortions we found that the multi-sensor method achieved the same or better hit rate than the full search with a fraction of running time. Also our approach makes it possible to propose optimal viewing angle for recognition as shown in Fig 9.

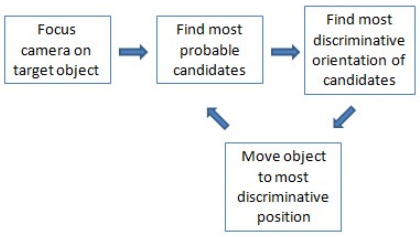

Fig. 9. Interactive recognition cycle.

Acknowledgment

The work of L. C. was supported by Bolyai scholarship of the Hungarian Academy of Sciences.

References

[1] O. Javed, M. Shah and D. Comaniciu. A probabilistic framework for object recognition in video. *In International Conference on Image Processing (ICIP)*, Page 2713-2716, 2004.

[2] A. Bruno, L. Greco and M. Cascia. Video Object Recognition and Modeling by SIFT Matching Optimization. *In ICPRAM* , page 662-670, 2014.

[3] J. He, J. Feng, X. Liu, T. Cheng, T. Lin, H. Chung and S. Chang. Mobile Product Search with Bag of Hash Bits and Boundary Reranking. *In IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2012.

[4] S. Gammeter, A. Gassmann, L. Bossard, T. Quack and L. Gool. Server-side object recognition and client-side object tracking for mobile augmented reality. *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 1-8, 2010.

[5] L. Czuni, Kiss P. J., A. Lipovits, M. Gal. Lightweight mobile object recognition. *In IEEE International Conference on Image Processing (ICIP)*, page 3426-3428, 2014.

[6] A. Pernek, L. Hajder and C. Kao. Metric Reconstruction with Missing Data under Weak Perspective. *In British Machine Vision Conference*, page 685-694, 2008.

[7] S. A. Chatzichristofis and Y. S. Boutalis. Accurate Image Retrieval based on Compact Composite Descriptors and Relevance Feedback Information. *International Journal of Pattern Recognition and Artificial Intelligence*, page 207-244, 2010.

[8] T. Sikora. The MPEG-7 visual standard for content description-an overview. *In IEEE Transactions on Circuits and Systems for Video Technology*, vol.11, no.6, page 696-702, 2001.

[9] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *In Int. Journal of Computer Vision*, page 91-110, 2004.

[10] O. Miksik and K. Mikolajczyk. Evaluation of local detectors and descriptors for fast feature matching. *In 21st International Conference on Pattern Recognition (ICPR)*, page 2681-2684, 2012.

[11] Z. Chi, H. Yan and T. Pham. Fuzzy Algorithms: With Applications to image processing and pattern recognition. In Advances in Fuzzy Systems - Applications and Theory, Vol.10, 1996.

[12] C. Rother, V. Kolmogorov and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *In ACM Transactions on Graphics (TOG)*, page 309-314, 2004.

[13] A. R.J. Francois. CAMSHIFT tracker design experiments with Intel OpenCV and sai. *University OF Southern California Los Angeles Inst. for Robotics and Intelligent Systems*, 2004.

[14] H. H. Bulthoff, C. Wallraven and A. Graf. View-based dynamic object recognition based on human perception. *16th International Conference on Pattern Recognition*, 2002.